AD_____

Award Number: DAMD17-03-1-0015

TITLE: A Likelihood Ratio Classifier for Computer-Aided Diagnosis in Mammography

PRINCIPAL INVESTIGATOR: Anna Bilska-Wolak

CONTRACTING ORGANIZATION: Duke University
Durham, NC 27705

REPORT DATE: April 2006

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01-04-2006	Annual Summary	13 Jun 2003 – 13 Apr 2006

4. TITLE AND SUBTITLE

A Likelihood Ratio Classifier for Computer-Aided Diagnosis in Mammography

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-03-1-0015

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Anna Bilska-Wolak

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University
Durham, NC 27705

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains colored plates: ALL DTIC reproductions will be in black and white.

14. ABSTRACT

In this research we developed a highly sensitive and specific computer-aided diagnosis classifier based on the likelihood ratio (LRb). The classifier is designed to aid physicians to identify mammographic lesions that should not be sent to biopsy. The classifier was developed using a large database of over five thousand breast biopsy cases from several medical centers. As a result of our research, we have developed a likelihood ratio classifier that can predict biopsy outcome for mass lesions. The performance of the classifier has been tested rigorously including testing on data that was not used for training, and also on data that originated from different medical centers. The results suggest that the LRb is a robust classifier for prediction of biopsy outcome. By decreasing the number of benign mass cases sent to biopsy, the classifier could be a valuable tool for physicians and ultimately beneficial to hospitals and patients.

15. SUBJECT TERMS
computer-aided diagnosis, mammography, likelihood ratio, biopsy, case-based reasoning, artificial intelligence

16. SECURITY CLASSIFICATION OF:				18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					USAMRMC
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	25	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

INTRODUCTION

Although screening x-ray mammography has become a very sensitive method for detecting breast cancer, mammography has low specificity in its diagnostic stage. About 67-85% of breast biopsies are performed on benign lesions. Because of cost and detrimental effects of unnecessary biopsies, the number of biopsies performed on benign lesions needs to be reduced. In this research we developed a highly sensitive and specific computer-aided diagnosis classifier based on the likelihood ratio, which is designed to aid physicians to identify lesions that should not be sent to biopsy. The classifier was developed using a large database of over five thousand breast biopsy cases from several medical centers. The cases present in the databases were described using BI-RADS™ lexicon and patient history, and represent the collective knowledge of physicians. The resulting classifier is statistically based, mathematically simple, and computationally efficient. Rigorous and exhaustive classifier evaluation methods included Receiver Operating Characteristic (ROC) analysis and leave-one-out bootstrap sampling.

STATEMENT OF WORK (01-2004)

Task 1. Develop and optimize case representation and database of over 4500 biopsy cases. (Months 1-36)

a. Previously acquired cases from Duke University, University of Pennsylvania (Penn)

b. Continue extracting information for cases from the (DDSM) database of University of South Florida

c. Acquire cases from other medical institutions

Task 2. Develop the LR and optimize its subcomponents. (Months 1-24)

a. Optimize mathematical feature representation from categorical case data

b. Estimate and optimize the N-dimensional density distribution of features (histogram approach, histogramming with smoothing functions, nearest-neighbor approaches, optimal decision fusion, kernel-density estimation)

c. Optimize features used (exhaustive search techniques, singular value decomposition, principle component analysis)

d. Evaluate model using ROC analysis, Round Robin sampling, and bootstrap

Task 3. Evaluate the performance of the LR under various conditions stemming from the input data. (Months 12-30)

a. Train and test separately on data from different institutions (i.e. train on all cases, test on cases from Duke University)

b. Train and test separately on different lesion types (i.e. train on all cases, test on mass lesions)

Task 4. Simulate and evaluate the use of LR in a clinical setting. (Months 24-36)

a. Analyze the optimized classifier on a set of data not used in training/development.

b. Examine how the standards (sensitivity, specificity) set on the training data affect the sensitivity and specificity on the new test data

c. Establish guidelines for retraining the classifier when a significant amount of new data is added

d. Conduct a retrospective clinical evaluation to evaluate LR's influence on physician's performance.

BODY

Task 1. Develop and optimize case representation and database of over 4500 biopsy cases. (Months 1-36)

- The initial hypothesis was verified experimentally.
- A large database of cases has been obtained and adapted for the project.
- The initial database was subsequently increased by approximately 400 additional cases from Duke University, 400 cases from Sloan-Kettering Cancer Institute, 350 cases from the University of North Carolina, and 125 cases from University of Maryland.
- This task has been completed and surpassed as proposed in the Statement of Work (100% completed).

Task 1 was an effort to collect as many cases as possible to aid in developing our classifier. The initial data were augmented by two new datasets from University of North Carolina at Chapel Hill (UNC) and from University of Maryland. Please refer to Annual Report 1 for more specifics.

This year, data from University of North Carolina at Chapel Hill were cleaned up and converted into a format consistent with the rest of our data. This was a major task, since the encoding for the cases was quite different from ours. Also, the cases included extra features and palpable cases, which we could not use. After the cleanup and conversions, there were 260 new available cases from UNC.

Data that were collected at the University of Maryland proved impossible to transcribe into the format that was used in the project. We were unable to obtain a legend for this dataset of 125 cases. Foremost, biopsy outcome was not clearly delineated and a strange numbering scheme was used. Subsequently, we were unable to use the Maryland data in our analysis.

In summary, data sets utilized in the project included data from five institutions: Duke University Medical Center (Duke); University of Pennsylvania Medical Center (Penn); Sloan Kettering Memorial Cancer Center (SK); public-database data collected by the University of South Florida (USF); University of North Carolina at Chapel Hill (UNC). After removing inconsistencies and palpable cases, 5561 cases in total were available for classifier development. This task has been completed and outdone with respect to the Statement of Work.

Task 2. Develop the LR and optimize its subcomponents. (Months 1-24)

- A likelihood ratio (LRb) classifier has been implemented.
- The mathematical feature representation has been optimized for the available data and classifier.
- The N-dimensional density distribution of features has been optimized using the nearest-neighbor approach, histogram approach, multivariate normal assumption, and decision fusion.
- Features used have been optimized using optimal exhaustive search techniques.
- Use of ROC analysis, Round Robin sampling and bootstrap were under consistent utilization for each classifier version.
- This task is 100% completed.

For **Task 2**, a likelihood ratio based classifier (LRb) has been developed, and extensive optimization of its subcomponents has been performed.

Task 2a. We evaluated the following approaches for BI-RADS™ feature optimization: 1) numerical rankings; and 2) histogram-based. We have concluded that the best feature representation for our features is one that is independent of ranking scales, and follows naturally from the data presented - the categorical histogram approach. Our data supports this conclusion, since best performance had been achieved with the histogram-encoded version of the classifier. Please consult Annual Report 1 for specifics on Task 2a.

For **Task 2b** we estimated and optimized feature density distributions. We completed nearest-neighbor approaches and histogram approaches; refer to Annual Report 1. We have also performed density estimation using the multivariate normal assumption and decision fusion. Please refer to Annual Report 2.

For feature optimization (**Task 2c**), nearest neighbor approach exhaustive search technique has been completed. Please see Annual Report 1. Singular value decomposition and principle component analysis results were completed as follows.

Principle Component Analysis

Principle component analysis (PCA) is an unsupervised linear feature extraction method.[1, 2] PCA computes the K largest eigenvalues of the covariance matrix of the d-dimensional pattern space.[2] These K largest eigenvalues can imply a natural dimensionality of the (transformed) data.[1] However, since PCA transforms the original features, the eigenvalues correspond to the transformed space and therefore the K transformed features need to be used as input to the classifier. PCA was performed on the Duke and Penn subsets. The features were transformed using PCA and then the top principal components based on variance were used as input to a Gaussian-based linear classifier. We could not use the categorical ranking approach, since the transformed features were continuous. For the Duke set, using a leave-one-out evaluation, the ROC area under the curve was 0.90, and the 0.90AUC was 0.56. While these values did not differ from the ROC results from other evaluations, a clinically significant difference was seen in the areas of high sensitivity. At 100% sensitivity, only 5% benign cases would be spared from biopsy with the PCA approach, as compared to the LRb. Similarly for the Penn dataset, the AUC was 0.85, and PAI was 0.35, while 0% benign cases would be spared at 100% sensitivity. While the PCA was a strong approach resulting in a fast linear classifier that produced high AUCs, the performance was not satisfactory in the high sensitivity regions.

Since PCA and singular value decomposition are related, no merit was seen in performing singular value decomposition after poorer results from PCA.

As specified in **Task 2d**, ROC analysis, Round Robin sampling and bootstrap have been under consistent utilization for all evaluations in Tasks 2-4. All model versions were evaluated and re-evaluated with ROC and various sampling methods throughout the duration of the research.

Task 3. Evaluate the performance of the LR under various conditions stemming from the input data. (Months 12-30)

- Training and testing of the LRb classifier has been performed on cases from all medical centers.
- Training and testing of the LRb classifier has been performed on different lesion types (masses vs. calcifications).
- This task is 100% completed.

Task 3a and 3b. In order to evaluate the performance of the LRb classifier on the various subsets of our large multi-institutional database, we performed a very comprehensive evaluation.

In Annual Report 1, evaluations were carried out on a) a subset of cases from two institutions, Duke University and University of Pennsylvania, and b) on different lesions types from Duke University data. Here we report on the rest of the multi-institutional evaluation for all .

Cases with Mass Lesions – Cross-Institutional Evaluation

All cases containing masses were extracted from each dataset. Mass cases were defined as containing a mass and any other findings. The cases were described using 16 features based on the BI-RADS lexicon and patient history. The features are listed in Annual Report 1. The mass datasets and the number of cases in each are listed in Table 1. The number of cases in the datasets ranged from 96 (UNC set) to 1196 (USF dataset). The percentage of malignant cases in each data set ranged from 28% to 53%, with an average of 40% malignancies per dataset.

Table 1: Number of **mass** cases for each Medical Center. The last column shows the number of benign cases that were actually sent to biopsy.

Institution Name/Dataset	Number of Mass Cases	Num. Malignant	Percent Malignant (%)	Num. Benign	Percent Benign (%)
Duke	670	244	36%	426	64%
Duke Set 2	151	43	28%	108	72%
SK	171	90	53%	81	47%
UPenn	496	200	40%	296	60%
UNC	96	29	30%	67	70%
USF	1196	615	51%	581	49%
Sum	2780	1221		1559	
Average	463	204	40%	260	60%

For this evaluation experiment, all features were used as input to the categorical-histogram LRb classifier. Round Robin evaluations were first carried out one each set separately, with bootstrap applied to the outputs to evaluate variance. The results of this evaluation are listed in Table 2 (column 2 and 3). Also, leave-one-out bootstrap (l-v-bootstrap) was applied to each set. This means that for each l-v-bootstrap evaluation, the LRb classifier was trained and tested 3000 times, and the final AUC and PAI values were averaged over the 3000 samples. The l-v bootstrap performance metrics generalize better to a population of testers and trainers. These l-v-bootstrap results are listed in Table 2, columns 4 and 5.

The best self-performance using the l-v-bootstrap was on our original Duke data set. These results appear consistent, since more optimization time was spent on the Duke set, and also the Duke dataset had values for most of

the features, in contrast to the other datasets. The AUC was 0.90+/-0.02 and the PAI was 0.60. The next best performance was UPenn, followed by USF. The three smallest dataset (Duke Set 2, SK, and UNC) had the lowest performance in terms of AUC (0.85,0.83,0.84). This suggests that they had the least information to predict on themselves, and not enough information was present to adequately populate the feature space.

Table 2: Training Results on Each Institution Separately for Mass Lesions

Institution Name/Dataset	Leave-one-out AUC	Leave-one-out PAI	Leave-one-out bootstrap AUC	Leave-one-out bootstrap PAI
Duke	0.90+-0.01	0.61+-0.04	0.90+-0.02	0.60+-0.06
Duke Set 2	0.86+-0.03	0.43+-0.17	0.85+-0.05	0.41+-0.22
SK	0.83+-0.03	0.52+-0.09	0.83+-0.05	0.51+-0.12
UPenn	0.88+-0.02	0.43+-0.07	0.88+-0.02	0.44+-0.09
UNC	0.84+-0.04	0.57+-0.09	0.84+-0.06	0.53+-0.17
USF	0.87+-0.01	0.49+-0.03	0.87+-0.01	0.50+-0.05
Average	0.86+-0.02	0.51+-0.08	0.86+-0.04	0.50+-0.12

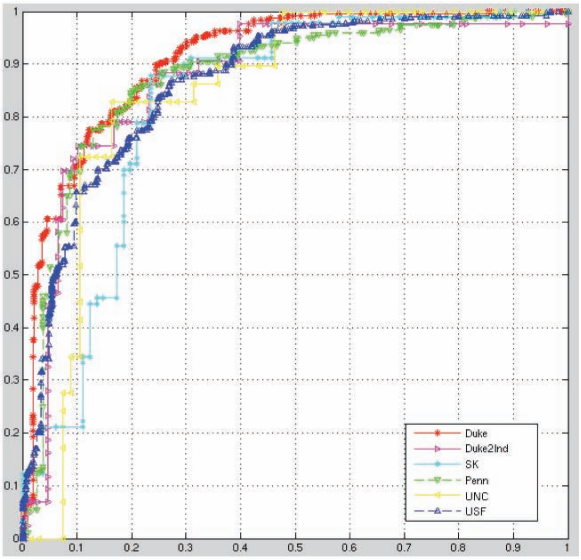

Figure 1: Training ROC curves for each mass cases from each institution.

The comprehensive results of the multi-institutional evaluation for the mass cases from all the medical centers are listed in Table 4. All possible institutional combinations were analyzed. For example, the Duke set was used as the training set for all the other data sets. The Duke was also used as the testing set for all the other data sets.

Testing on Duke

Figure 2A shows the ROC curves produced when the LRb classifier was trained on the various datasets, and tested on the Duke set. Training with dataset with the smallest number of cases (SK, UNC, Duke Set 2) produced smallest AUC

areas (0.87,0.85,0.84). The average AUC value for testing on the Duke set was 0.87+/-0.02. The average PAI was 0.43+-0.05.

Testing on Duke Set 2

Figure 2B shows the ROC curves produced when the LRb classifier was trained on the various datasets, and tested on the Duke Set 2. The average AUC value for testing on the Duke Set 2 was 0.84+/-0.04. The average PAI was 0.44+-0.11.

A very small performance was seen when the training data for this set was the UNC set. The area was 0.75+/-0.03. This is probably due to the fact that the two sets had a lot of missing feature values that differed between the sets.

Table 3: Cross Training and Testing on Various Medical Centers: AUC results for mass lesions. The first left column represents the Training sets. The top row represents the Test sets.

	Test						
Train: Institution Name/Dataset	Duke	Duke Set 2	SK	UPenn	UNC	USF	Average
Duke	-	0.88+-0.03	0.83+-0.03	0.88+-0.02	0.88+-0.04	0.86+-0.01	0.87+-0.03
Duke Set 2	0.84+-0.02	-	0.83+-0.03	0.86+-0.02	0.86+-0.04	0.81+-0.01	0.84+-0.02
SK	0.87+-0.01	0.86+-0.03	-	0.88+-0.02	0.82+-0.05	0.85+-0.01	0.86+-0.03
UPenn	0.90+-0.01	0.83+-0.04	0.84+-0.03	-	0.85+-0.04	0.87+-0.01	0.86+-0.03
UNC	0.85+-0.02	0.75+-0.04	0.82+-0.03	0.88+-0.02	-	0.84+-0.01	0.83+-0.02
USF	0.89+-0.01	0.87+-0.03	0.84+-0.03	0.88+-0.02	0.84+-0.05	-	0.86+-0.03
Average	0.87+-0.01	0.84+-0.04	0.83+-0.03	0.88+-0.02	0.85+-0.04	0.85+-0.01	

Testing on SK

Figure 2C shows the ROC curves produced when the LRb classifier was trained on the various datasets, and tested on the SK set. The average AUC value for testing on the SK set was 0.83+/-0.02. The average PAI was 0.51+-0.06. The PAI performance on this set was most exceptional of all the sets, indicating that a lot of benign cases could be potentially spared from biopsy. This effect was consistent regardless of which dataset was used for training. The best SK performance was achieved when training on the Duke data set, with a PAI of 0.56+/-0.06.

Testing on UPenn

Figure 2D shows the ROC curves produced when the LRb classifier was trained on the various datasets, and tested on the UPenn set. The average value for testing on the UPenn set was 0.88+/-0.02 for AUC, and the average PAI was 0.41+-0.06. The AUC was consistently high at 0.88 for almost all training dataset.

Testing on UNC

Figure 2E shows the ROC curves produced when the LRb classifier was trained on the various datasets, and tested on the UNC set. The average AUC value for testing on the UNC set was 0.85+/-0.04 for AUC, and the average PAI was 0.42+-0.11. As for the other smaller datasets, the variances on these measurements were higher than for the larger datasets.

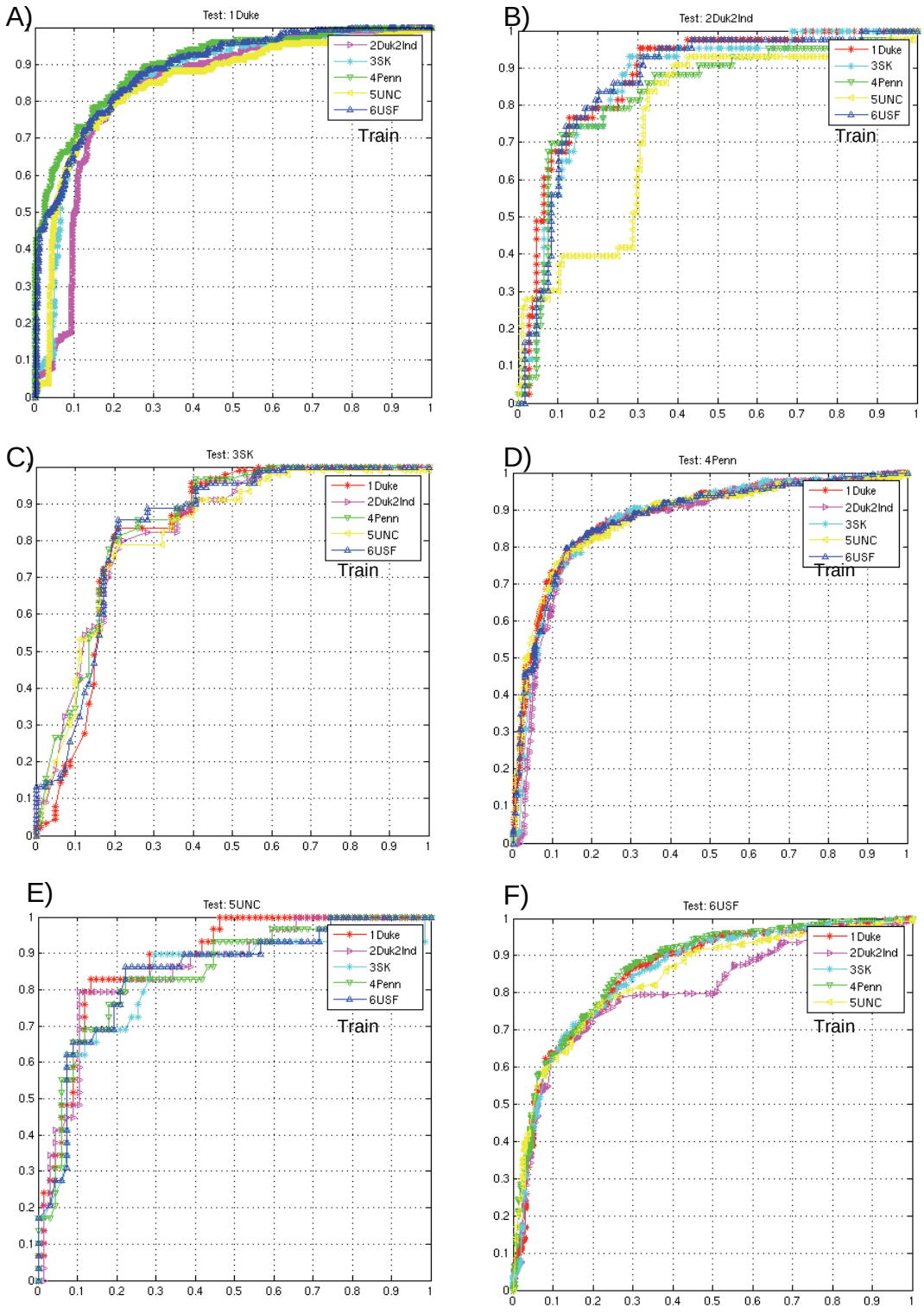

Figure 2: Round Robin ROC graphs for the cross-institutional evaluation of mass cases. For each graph, the x-axis is the FPF, and the y-axis is the TPF. A) ROC curves for testing on Duke dataset. B) Testing on Duke Independent Set 2. C) Testing on Sloan-Kettering. D) Testing on UPenn dataset. E) Testing on UNC dataset. F) Testing on USF dataset. You can see Figure 4 for the same ROC curves grouped by training data.

Testing on USF

Figure 2F shows the ROC curves produced when the LRb classifier was trained on the various datasets, and tested on the USF set. The average AUC value for testing on the UNC set was 0.87+/-0.01 for AUC, and the average PAI was 0.36+-0.03.

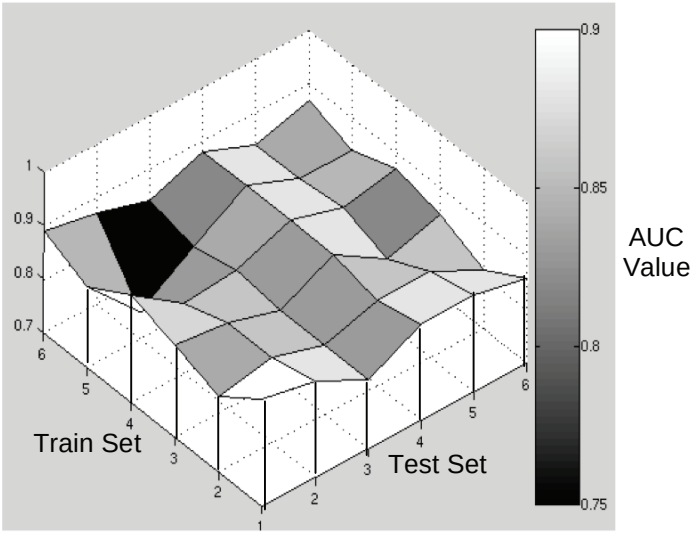

Figure 3: AUC values for the cross-institutional evaluation: results from Table 4 visualized as a surface plot. Duke = 1, Dukeset2 = 2, SK =3, UPenn =4, UNC=5, USF=6. Empty spots in Table 4 were filled with the self-evaluation of each test from Table 3.

Table 4: Cross-training and testing on mass lesions from various medical centers: PAI results

	Test						
Institution Name/Dataset	Duke	Duke Set 2	SK	UPenn	UNC	USF	Average
Duke	-	0.58+-0.11	0.56+-0.06	0.40+-0.07	0.59+-0.09	0.40+-0.04	0.51+-0.06
Duke Set 2	0.41+-0.05	-	0.49+-0.06	0.39+-0.07	0.45+-0.11	0.23+-0.03	0.39+-0.05
SK	0.45+-0.05	0.52+-0.12	-	0.44+-0.07	0.24+-0.21	0.41+-0.04	0.41+-0.08
UPenn	0.51+-0.05	0.32+-0.14	0.55+-0.06	-	0.42+-0.12	0.45+-0.04	0.45+-0.07
UNC	0.31+-0.06	0.25+-0.17	0.42+-0.08	0.38+-0.08	-	0.31+-0.04	0.33+-0.07
USF	0.49+-0.05	0.53+-0.14	0.53+-0.07	0.41+-0.07	0.39+-0.14	-	0.47+-0.08
Average	0.43+-0.04	0.44+-0.11	0.51+-0.06	0.41+-0.06	0.42+-0.11	0.36+-0.03	

Train

It is interesting to note that for the small datasets, the best performance was achieved when a larger dataset was used for training. Specifically, for the SK, UNC, and Duke Set 2 (which are the smallest sets) the best performance was achieved when either the Duke set or the Penn set were used for training. For the large datasets (Duke, UPenn, USF), the performance for

training/testing on themselves was just as high as any of the cross-institutional evaluations.

The results of the multi-institutional evaluation demonstrate the importance of 1) having an adequate number of cases to train the classifier, and 2) having cases with as many feature values filled as possible. As evident from evaluations on the smaller datasets, too few cases and too few features have a detrimental effect on performance.

The consistently high AUC and PAI results across institutional evaluations over all demonstrate that the classifier is able to predict on data from different medical centers. This also shows that it is possible to create a classifier for breast biopsy prediction that works across medical institutions.

Figure 4 shows the same ROC curves as Figure 3, however grouped by training set. As it is evident by comparing Figure 3 and 5, the same of the ROC curve is driven more by the testing set, than by the training set.

The last evaluation that was carried out is shown in Figure 5. All sets except the testing set were used as the training data. For example, when UNC was used as the testing set, the rest of the sets (Duke, Duke Set 2, SK, UPenn, and USF) were used as training data. The performance on this evaluation was the best of all, showing that it is best to train using all available data.

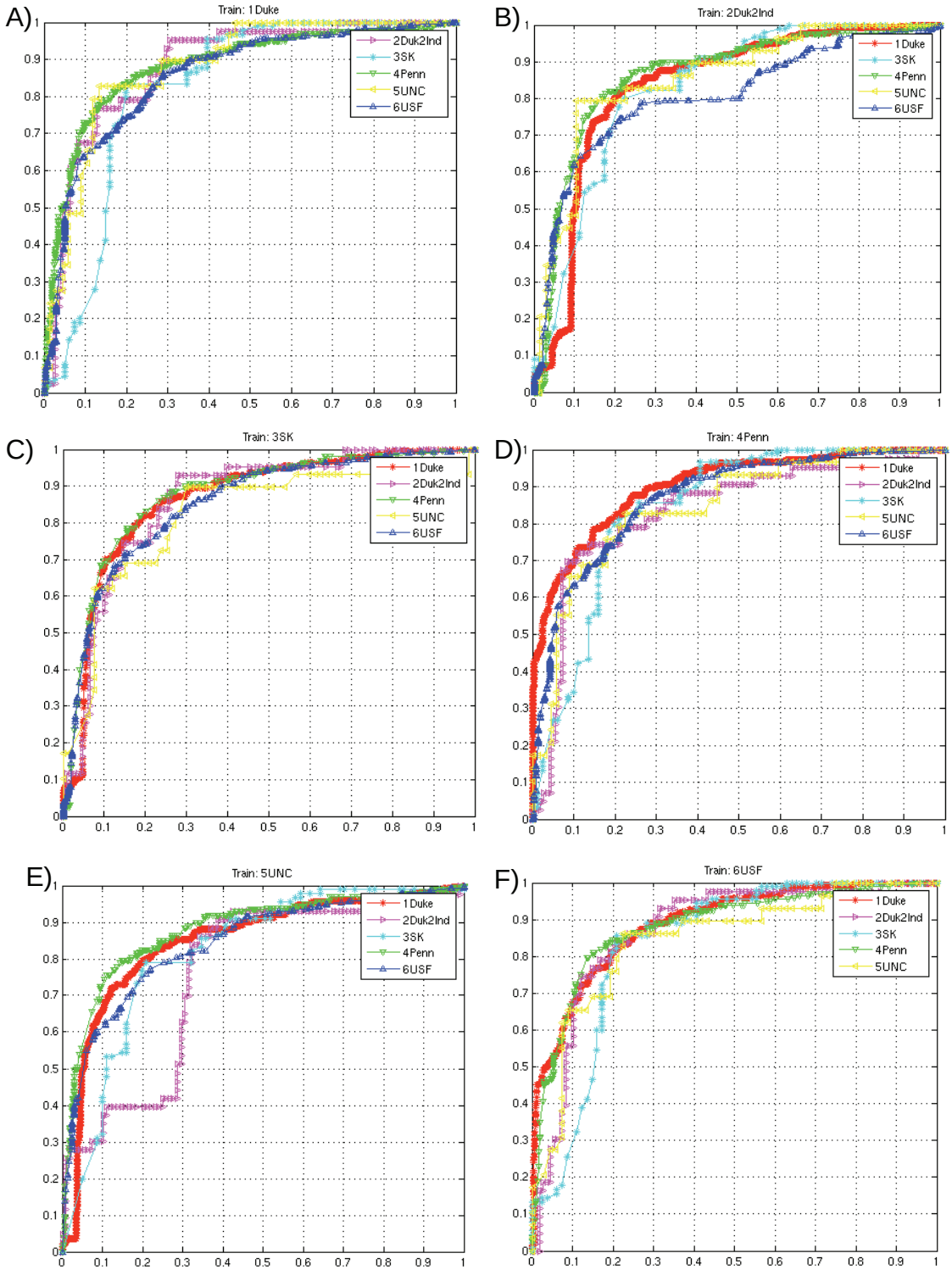

Figure 4: Round Robin ROC graphs for the cross-institutional evaluation of mass cases. This figure shows the same ROC curves as in Figure 2, but grouped by the training dataset. For each graph, the x-axis is the FPF, and the y-axis is the TPF. A) ROC curves for training on Duke dataset. B) Training on Duke Independent Set 2. C) Training on Sloan-Kettering. D) Training on UPenn dataset. E) Training on UNC dataset. F) Training on USF dataset.

	AUC	PAI
Duke	0.89+-0.01	0.50+-0.01
Duke Set 2	0.89+-0.03	0.65+-0.03
SK	0.83+-0.03	0.54+-0.03
UPenn	0.89+-0.02	0.44+-0.02
UNC	0.88+-0.04	0.58+-0.04
USF	0.86+-0.01	0.42+-0.01

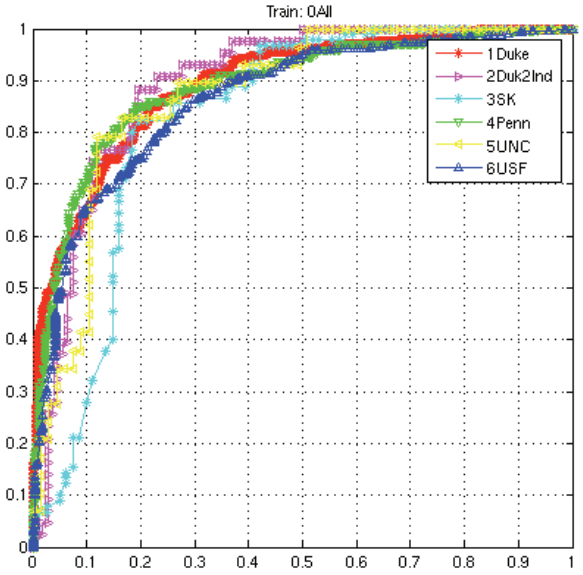

Figure 5: The ROC curves for all datasets when trained on the full database. For example, the red curve shows the performance of the classifier when the Duke2Ind, SK, Penn, UNC and USF datasets were used for training, and the Duke set was used for testing.

Cases with calcification lesions – Cross-institutional Evaluation

Table 5: Number of cases with calcification lesions for each institution.

Institution Name/Dataset	Number of Calc Cases	Num. Malignant	Percent Malignant(%)	Num. Benign	Percent Benign (%)
Duke	671	224	33%	447	67%
Duke Set 2	223	64	29%	159	71%
SK	310	129	42%	181	58%
UPenn	478	190	40%	288	60%
UNC	151	55	36%	96	64%
USF	768	360	47%	408	53%
Sum	2601	1022		1579	
Average	434	170	38%	451	62%

All cases containing calcifications were extracted from each dataset. Calcification cases were defined as containing calcifications and any other findings, but no masses. The cases were described using 16 features based on the BI-RADS lexicon and patient history. The features are listed in Annual Report 1. The datasets and the number of calcification cases in each are listed in Table 7. The number of cases in the datasets ranged from 151 (UNC set) to 768 (USF dataset). The percentage of malignant cases in each data set ranged from 29% to 47%, with an average of 38% malignancies per dataset.

As with our previous evaluations, we were unable to obtain clinically satisfactory performance on identifying likely-benign calcification cases. The ROC areas ranged from 0.56 (practically chance performance) to 0.77, as shown in Table 6 and Figure 6. This suggests that our BI-RADS findings did not contain enough information to predict on calcification cases. The classifier should not be applied to classification of calcifications.

Table 6: Training results for calcification cases for institution trained and tested on itself in a leave-out-out (Round Robin) fashion.

Institution Name/Dataset	Leave-one-out AUC	Leave-one-out PAI	Leave-one-out bootstrap AUC	Leave-one-out bootstrap PAI
Duke	0.66+-0.02	0.12+-0.03	0.65+-0.03	0.12+-0.04
Duke Set 2	0.56+-0.05	0.01+-0.02	0.55+-0.06	0.04+-0.04
SK	0.60+-0.03	0.09+-0.02	0.63+-0.04	0.11+-0.04
UPenn	0.77+-0.02	0.22+-0.05	0.77+-0.03	0.22+-0.06
UNC	0.67+-0.05	0.15+-0.06	0.66+-0.07	0.15+-0.09
USF	0.68+-0.02	0.13+-0.02	0.69+-0.03	0.13+-0.03
Average	0.66+-0.03	0.12+-0.03	0.66+-0.04	0.13+-0.05

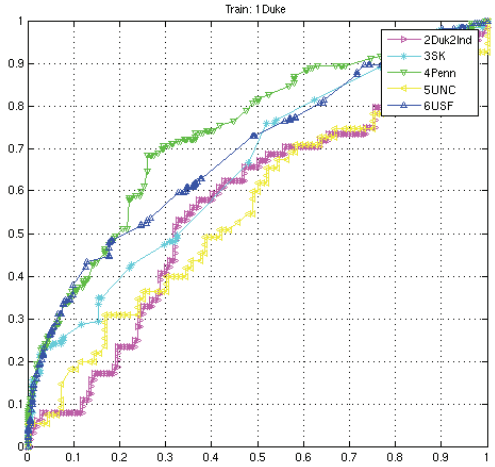

Figure 6: ROC areas for calcifications cases when Duke was used as the training data, and all other data sets were used for testing

Cases with all lesion types – Cross-institutional Evaluation

For this evaluation, all lesion types were used, including masses, calcifications and special cases. The datasets and the number of cases in each are listed in Table 1. The number of cases in the datasets ranged from 260 (UNC set) to 1979 (USF dataset). The percentage of malignant cases in each data set ranged from 27% to 50%, with an average of 39% malignancies per dataset.

Table 7: Number of cases of all lesion types for each institution.

Institution Name/Dataset	Number of Cases	Num. Malignant	Percent Malignant (%)	Num. Benign	Percent Benign (%)
Duke	1497	521	35%	976	65%
Duke Set 2	444	122	27%	322	73%

SK	481	219	46%	262	54%
UPenn	1000	396	40%	604	60%
UNC	260	91	35%	169	65%
USF	1979	985	50%	994	50%
Sum	5661	2334		3327	
Average	944	389	39%	555	61%

The classifier results for all lesions were obviously affected by the presence of calcification cases, which we have already determined to have poor performance.

Table 8: Training Results on Each Institution Separately – all lesion types. For the leave-one-out bootstrap, the classifier was re-trained on different subsets of the given medical center data 3000 times.

Institution Name/Dataset	Leave-one-out AUC	Leave-one-out PAI	Leave-one-out bootstrap AUC	Leave-one-out bootstrap PAI
Duke	0.81+-0.01	0.33+-0.03	0.80+-0.02	0.31+-0.05
Duke Set 2	0.69+-0.03	0.15+-0.04	0.67+-0.05	0.14+-0.07
SK	0.70+-0.02	0.13+-0.03	0.71+-0.04	0.15+-0.05
UPenn	0.83+-0.01	0.32+-0.04	0.83+-0.02	0.31+-0.06
UNC	0.77+-0.03	0.33+-0.06	0.74+-0.05	0.27+-0.09
USF	0.81+-0.01	0.34+-0.02	0.81+-0.01	0.34+-0.03
Average	0.77+-0.02	0.27+-0.04	0.76+-0.03	0.25+-0.06

Table 8: Cross training and testing on various medical centers and lesion types: AUC results

	Test						
Train: Institution Name/Dataset	Duke	Duke Set 2	SK	UPenn	UNC	USF	Average
Duke	-	0.72+-0.03	0.74+-0.02	0.81+-0.01	0.67+-0.03	0.80+-0.01	0.75+-0.02
Duke Set 2	0.73+-0.01	-	0.71+-0.02	0.79+-0.01	0.71+-0.03	0.73+-0.01	0.73+-0.02
SK	0.75+-0.01	0.70+-0.03	-	0.78+-0.02	0.67+-0.04	0.79+-0.01	0.74+-0.02
UPenn	0.79+-0.01	0.70+-0.03	0.75+-0.02	-	0.71+-0.03	0.80+-0.01	0.75+-0.02
UNC	0.65+-0.01	0.64+-0.03	0.67+-0.03	0.72+-0.02	-	0.70+-0.01	0.68+-0.02
USF	0.78+-0.01	0.70+-0.03	0.74+-0.02	0.80+-0.01	0.71+-0.03	-	0.75+-0.02
Average	0.74+-0.01	0.69+-0.03	0.72+-0.02	0.78+-0.01	0.69+-0.03	0.01	

Table 9: Cross training and testing on various medical centers and all lesion types: PAI results

Train \ Test	Duke	Duke Set 2	SK	UPenn	UNC	USF	Average
Duke	-	0.24+-0.04	0.25+-0.04	0.31+-0.04	0.18+-0.04	0.30+-0.02	0.25+-0.03
Duke Set 2	0.21+-0.02	-	0.17+-0.03	0.26+-0.03	0.21+-0.04	0.17+-0.02	0.20+-0.02
SK	0.18+-0.02	0.19+-0.06	-	0.22+-0.03	0.12+-0.05	0.22+-0.02	0.18+-0.03
UPenn	0.32+-0.03	0.19+-0.04	0.24+-0.03	-	0.15+-0.04	0.33+-0.02	0.25+-0.03
UNC	0.16+-0.02	0.12+-0.04	0.16+-0.03	0.24+-0.03	-	0.18+-0.02	0.17+-0.02
USF	0.24+-0.02	0.12+-0.05	0.22+-0.04	0.28+-0.03	0.16+-0.04	-	0.21+-0.03
Average	0.22+-0.02	0.17+-0.04	0.21+-0.03	0.26+-0.03	0.16+-0.04	0.24+-0.02	

Based on the comparative results between lesion types, it is best to treat mass cases, calcification cases and other lesion types separately.

Task 4. Simulate and evaluate the use of LR in a clinical setting. (Months 24-36)

- Three independent evaluations on a set of mass cases previously unseen by the LRb classifier have been carried out.
- Standards of sensitivity and specificity have been evaluated on the new data.
- This task is 70% completed.

Task 4. Three independent evaluations have been completed using a subset of the newly acquired cases.

Task 4a. The first evaluation consisted of evaluating the LRb classifier on cases that were previously unseen by the classifier, but originated from the same medical center. This evaluation was summarized in Annual Report 1 and updated in Annual Report 2.

In a second evaluation, validation of the LRb was carried out on a new independent database of cases that: 1) were not used for computer-aid development; and 2) originated from another medical institution – Sloan Kettering Center (SK). This evaluation was reported partially in Annual Report 2. Here we report on the translation of thresholds for the SK independent evaluation (**Task 4b**). It is important to note that this evaluation was carried out before the full evaluation in Task 3, and was thus still independent.

Our thresholds analysis was made more difficult than usual by the inclusion of cases with missing values. This is not a problem for a few missing values. However, some of are datasets do not have values for as many as 10 features, and this was the case for the SK dataset. While this lack of features may not affect the ROC curves, it does affect detrimentally the threshold translations. Ie, when thresholds are determined using 16 features, and then evaluated on 6, the thresholds will be too high on the new data, and with a final effect of not sparing any benign biopsies. For this evaluation therefore, assuming unknown disease prevalence, we used only 6 features for evaluating the thresholds. Four thresholds (established on the test data) were evaluated on the training set at the four sensitivities; 95%, 98%, 99%, and 100%. Table 11 shows how the sensitivity established on the training data affected the sensitivity on the validation set. The most conservative threshold used was 100% sensitivity. For the RR method, applying the 100% sensitivity threshold to the new dataset yielded 100% sensitivity and 11% specificity. This means that 11% of the benign lesions could be spared from biopsy. For the BB method, applying the 100% sensitivity threshold to the new dataset actually yielded 100% sensitivity, and specificity of 25%. This means that 25% of the benign masses would be spared from biopsy. All of the other thresholds (99-95%) performed conservatively, yielding 100% sensitivity on the new data set.

Table 10: Characteristics of the datasets used in the second independent evaluation.

Characteristic	Training/Testing	Independent Validation
Medical Center	Duke	Sloan-Kettering
# of Mass Cases	670	171
# Malignant	244 (36%)	90 (53%)
# Benign	426 (64%)	81 (47%)
Avg. Age (range)	56 years (24-87)	57 years (33-92)
# Mass + Calc	79 (12%)	13 (8%)
# Mass + Arch.Dist.	2	8

Table 11: Performance of threshold method RR and BB at 100-95% sensitivity levels as applied to the 171 cases from the Sloan-Kettering data set.

ROC 171 Original Sensitivity	Original Specificity (Max achievable)	BB Method Resulting Sensitivity	BB Method Resulting Specificity	RR Method Resulting Sensitivity	RR Method Resulting Specificity
100%	60%	100%	25%	100%	11%
99%	51%	100%	25%	100%	25%
98%	44%	100%	26%	100%	28%
95%	43%	100%	37%	100%	40%

In general, the sensitivity/threshold established on the training data results in lower sensitivity on the testing data. However, in our evaluations, the sensitivity was maintained at a high level. The results are encouraging for applying the classifier to new unknown data in a clinical setting.

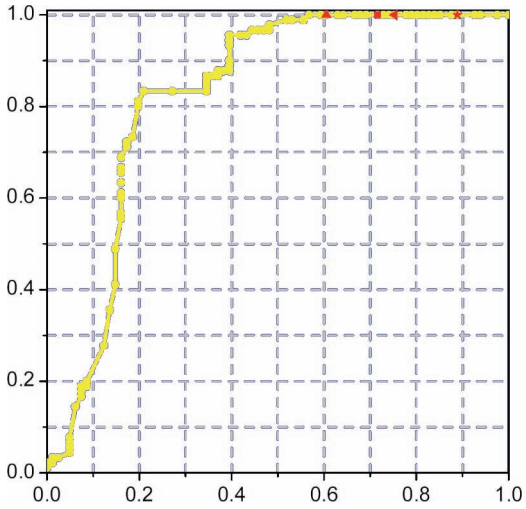

Figure 7: ROC curve for training on Duke set, testing on SK set. Thresholds established on the Duke set were applied to the SK set, and are marked in red.

Independent Evaluation - UNC data

The third evaluation was also a validation of the LRb on a new independent database of cases. These cases were not used for development and originated from the University of North Carolina at Chapel Hill. The data in this set consisted of several cases gathered at Good Samaritan, Massachusetts, Mt. Sinai, Toronto, UVA, Thomas Jefferson, and UNC Medical Centers.
The trained LRb classifier was applied to the UNC data, and the resulting ROC curve is plotted in Figure 8. The threshold analysis is included in Table 12.

Table 12: Performance of Thresholds from BB and RR methods as applied to the UNC set

Original ROC Sensitivity	Original ROC Specificity (Max Achievable)	BB Method Resulting Sensitivity	BB Method Resulting Specificity	RR Method Resulting Sensitivity	RR Method Resulting Specificity
95%	55%	83%	72%	83%	72%
98%	54%	90%	61%	90%	72%
99%	54%	100%	54%	93%	55%
100%	54%	100%	54%	100%	48%

Table 12 shows how the sensitivity established on the training data affected the sensitivity on the validation set. We used the four cutoff thresholds established from the training/test data (Duke set). The most conservative threshold used was 100% sensitivity. For the RR method, applying the 100% sensitivity threshold to the new dataset yielded 100% sensitivity and 48% specificity. This means that 48% of the benign lesions could be spared from biopsy. For the BB method, applying the 100% sensitivity threshold to the new dataset actually yielded 100% sensitivity, and specificity of 54%. This means that 54% of the benign masses would be spared from biopsy.

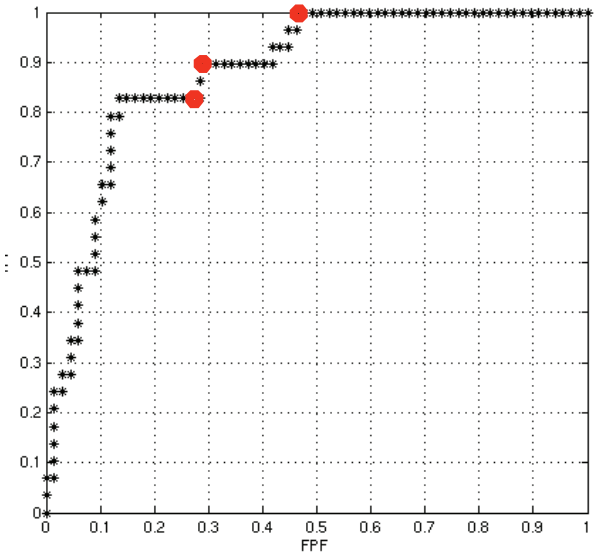

Figure 8: ROC curve obtained when the trained LRb classifier was applied to the UNC dataset. BB thresholds are plotted in red. The 100% sensitivity threshold was maintained, and would result in 54% specificity.

The performance on the classifier in three different independent evaluations has shown that the classifier can be translated to data from different medical centers. Even threshold translation was very commendable, maintaining high sensitivity and achieving good specificity on new unseen data. Our results indicate that the classifier could potentially spare benign cases from biopsy if it was used in the clinical setting.

Task 4d was not performed due to lack of time and resources. For evaluating the physician's performance with the classifier, a separate user interface (computer program) would have to be developed. This would allow the physician to interact with the classifier in a clinical-like setting.

ABBREVIATIONS

ACR	American College of Radiology
ACS	American Cancer Society
ANN	Artificial Neural Network
AUC	Area under the ROC curve
0.90AUC	Partial area under the ROC curve (Pd>0.90)
BI-RADS™	Breast imaging reporting and data system
BX	Biopsy outcome (malignant or benign)
CAD	Computer-Aided Diagnosis
CBLR	Case-Based Likelihood Ratio
CBR	Case-Based Reasoning
DDSM	Digital Database for Screening Mammography
Duke	Duke University Medical Center
FPF	False positive fraction
LRb	Likelihood ratio based classifier
PAI	Partial area index, same as 0.90AUC
UPenn	University of Pennsylvania Medical Center
Pd	Probability of detection
Pf	Probability of false alarm
ROC	Receiver Operating Characteristic
SK	Sloan-Kettering Cancer Institute
TPF	True positive fraction
UNC	University of North Carolina at Chapel Hill
USF	University of South Florida
UVA	University of Virginia

KEY RESEARCH ACCOMPLISHMENTS

- A large and unique database of biopsy cases for mammography has been collected and adapted for the project. The initial database was increased by approximately 400 additional cases from Duke-University, and 350 cases from Sloan-Kettering Cancer Institute. The final database contained over 5600 mammographic cases that were originally sent to biopsy, and originated from several medical centers.
- A likelihood ratio classifier has been implemented and optimized verifying the initial hypothesis. This novel classifier works extremely well in classifying categorical data.
- The resulting classifier was tested rigorously and exhaustively on all lesion types and all medical centers in the database. This evaluation was the largest of its kind to evaluate the cross-institutional performance on BI-RADS findings from multiple medical centers.
- The new methods were developed for establishing ROC thresholds for a given sensitivity of interest.
- Independent evaluations have been carried out that tested the developed classifier on new unseen data. The classifier was shown to be able to generalize and spare benign cases that it has not previously seen.

REPORTABLE OUTCOMES:

Over the course of the project, the PI participated in and published the following:

[1] A.O. Bilska-Wolak, C.E. Floyd Jr., Joseph Y. Lo, “Improved sensitivity for breast cancer classification using a case-based likelihood ratio.” Medical Image Perception Conference, 2003, Durham NC, September 2003.

[2] A.O. Bilska-Wolak, C.E. Floyd Jr, "Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer," Phys. Med. Biol. 49, September 2004, pp. 4219-4237.

[3] Bilska-Wolak, A. O. “Evaluation of a mammographic computer-aid on an independent database of cases,” oral presentation for Medical Physics Seminar, Department of Biomedical Engineering, Department of Radiology, Duke University, October 28th (2004).

[4] Bilska-Wolak, A. O., C. E. Floyd Jr., J. Y. Lo, (2005). "Computer Aid for Decision to Biopsy Breast Masses on Mammography: Validation on New Cases." Academic Radiology 12(6), 2005, pp. 671-680.

[5] Bilska-Wolak, A. O., C. E. Floyd Jr., J. Y. Lo, (2005), “A likelihood ratio classifier for computer-aided diagnosis in mammography.” Breast Cancer Era of Hope Conference, June 2005, Philadelphia, Pennsylvania.

[6] Bilska-Wolak, A. O., C. E. Floyd Jr., J. Y. Lo (2005), “Validation of a classifier for mammographic masses on new data from another medical institution.” Medical Image Perception Conference, MIPS September 2005.

Other accomplishments:

- Acquisition of Ph.D. degree by the P.I.
- A large database of BI-RADS^{TM} descriptions for mammography cases from several medical institutions.

- A new classifier model that performs very well on categorical data.
- Project/PI featured in the Era of Hope 2005 Press Release. http://cdmrp.army.mil/pubs/press/eoh2005_new_strategies.htm

It is anticipated that two more peer-reviewed manuscripts will be published using the numerous results from the comprehensive multi-institutional evaluation presented in this report.

PERSONNEL RECEIVING PAY FROM THE RESEARCH EFFORT
PI, Anna Bilska-Wolak

CONCLUSIONS:

A likelihood ratio classifier has been developed and optimized for breast biopsy classification. The performance of the classifier was comparable to or better than other classifiers previously developed for breast biopsy classification.

The classifier was cross-tested comprehensively on cases from various medical centers, showing that it was able to generalize well among medical centers. An independent validation test on 151 cases from the same medical center showed that the trained classifier was able to identify 26% of benign mass lesions that should not be sent to biopsy, while still correctly diagnosing 100% of malignancies. Two independent validations were also carried out on cases that came from a different medical center than the training data. In both evaluations, the trained classifier was able to identify 11%-54% of benign mass lesions that could be spared from biopsy, while maintaining 100% sensitivity to malignant lesions. The performance of the classifier was robust even with some missing case data, allowing full utilization of all the information present in the databases.

By decreasing the number of benign cases sent to biopsy, the classifier could be a valuable tool for physicians and ultimately beneficial to hospitals and patients.

REFERENCES:

1. Duda, R.O., P.E. Hart, and D.G. Stark, *Pattern Classification*. Second ed. 2001, New York: John Wiley & Sons.
2. Jain, A.K., R.P.W. Duin, and J. Mao, *Statistical pattern recognition: a review.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. **22**(1): p. 4-37.